Supplementary Material for DreamLIP

Kecheng Zheng [⋆,1,2], Yifei Zhang[⋆,3], Wei Wu[4], Fan Lu[4],
Shuailei Ma[6], Xin Jin[5], Wei Chen[1], and Yujun Shen[2]

[1] State Key Lab of CAD&CG, Zhejiang University
[2] Ant Group [3] Shanghai Jiao Tong University
[4] University of Science and Technology of China
[5] Eastern Institute of Technology [6] Northeastern University, China
{zkechengzk, xiaomabufei, shenyujun0302}@gmail.com,
qidouxiong619@sjtu.edu.cn, {lufan,wuvy}@mail.ustc.edu.cn, jinxin@eias.ac.cn

A Long Caption in Multi-modality Learning

In this section, we discuss various studies related to the generation of extended captions within the context of text-to-image (T2I) synthesis. Several notable works [1, 4, 5] have employed Multimodal Large Language Models (MLLM) to produce comprehensive and rich captions for T2I tasks. These approaches let models better capture and draw a picture by utilizing more intricate and accurate captions of scenes. Meanwhile, in language-image pretraining tasks, our objective is for elaborate captions to leverage real-world images more effectively, thereby endowing multimodal foundation models with some additional capabilities (*e.g.*, Vision-Language Compositionality in Appendix D).

B Sampling from Mixture Generated Long Captions

In the main paper, we mainly use the long caption from ShareGPT4V as the training text. However, different MLLMs may focus on different regions of real-world images due to their difference of training process and data. Thus, the generated long/short captions from MLLMs can help each other. Inspired by this straightforward idea, we merge all captions from different MLLMs together, and sample the sub-captions from the set of merged captions. As shown in Figure 1, we use three kinds of MLLMs to generate the long captions and short captions, and then sample the sub-captions from it as the input of text encoder.

As shown in Table 1, the number of sub-captions and tokens exceeds the result of using ShareGPT4V alone. Further, our ablation studies on the mixture generated long captions reveal that an increased number of sub-captions correlates with enhanced performance, surpassing the result achieved with long captions solely from ShareGPT4V. These findings suggest that various MLLMs capture distinct image regions, providing complementary information.

We also analyze the mixture generated long captions, characterized by a greater number of tokens and subcaptions, can capture the contents of an image

⋆ Equal contribution

more comprehensively than one kind of captions as shown in Figure 2. In future work, we aim to investigate the synergistic effects of integrating additional MLLMs.

Table 1: Ablation study of sampling number of sub-captions from mixture generated long captions. We use ViT-B/16 as image backbone. 'CLIP*' refers to enhancing CLIP with image data augmentation following SimCLR [6]

K	Text Retrieval		Image Retrieval		Classification	Segmentation
	Flickr30k R@1	MSCOCO R@1	Flickr30k R@1	MSCOCO R@1	ImageNet Acc.(%)	VOC-20 mIOU
CLIP	29.4	14.3	19.9	10.2	16.0	62.7
CLIP*	32.6	14.8	21.4	11.5	20.3	64.4
2	68.1	39.6	54.3	29.1	30.7	83.1
4	72.6	44.1	58.2	32.6	30.1	85.5
6	74.9	44.9	59.2	33.4	31.9	85.4
8	75.0	45.6	60.8	33.8	32.8	86.6
10	74.3	47.2	61.6	34.8	33.6	87.4
12	75.5	46.4	61.5	35.0	34.3	86.7
14	75.9	47.4	61.8	35.0	33.9	87.2
16	75.3	48.1	61.7	35.1	34.7	87.8
18	74.4	46.4	62.4	34.9	34.6	88.2

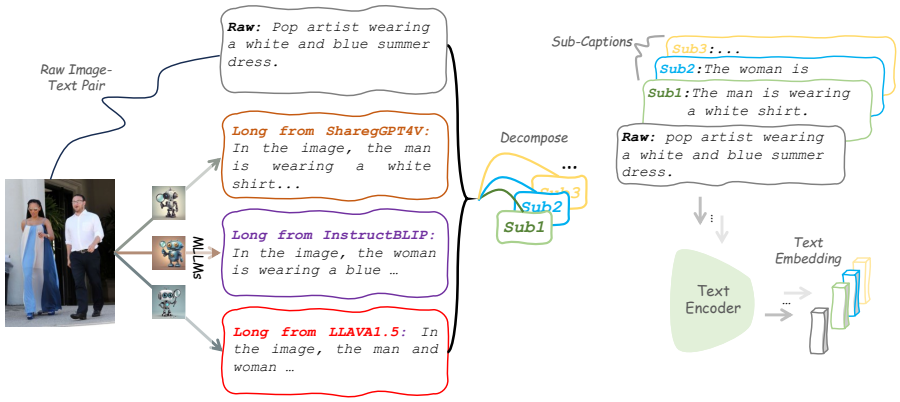


Fig. 1: Illustration of `DreamLIP` with mixture generated long captions.

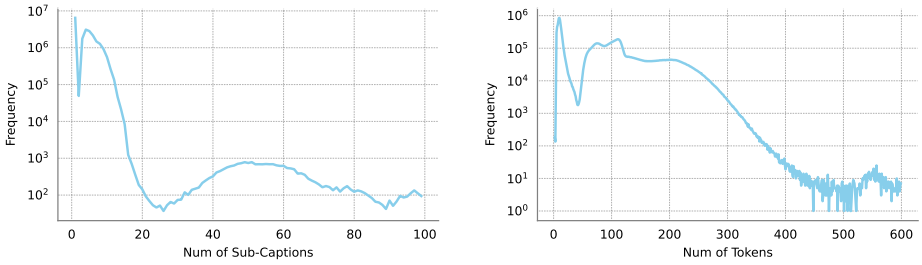


Fig. 2: Statistics of merged caption that include raw caption, three long captions and three short catptions generated by MLLMs (*i.e.*, InstructBLIP, LLAVA-1.5 and ShareGPT4V).

C Data Augmentation with Synthetic Images Generated by SDXL-turbo

Language-image pre-training could benefit from long captions generation, due to the strong captioning capacity of image-to-text models. Meanwhile, as shown in Figure 3, we would explore whether text-to-image models (*e.g.*, SDXL-turbo [9]) can bring performance improvement for language-image pre-training. An image may be depicted through a variety of sentences, while a single caption also has the capacity to evoke numerous images. Thus, we adopt the SDXL-turbo to generate some images from raw captions and short captions. As shown in Table 2, the DreamLIP with synthetic images outperforms DreamLIP on three kinds of datasets, which approves the ability of the introduction of synthetic images.

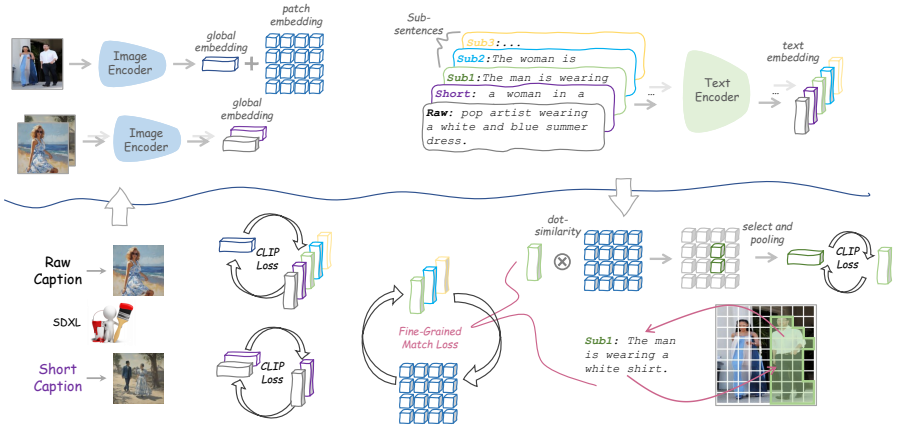


Fig. 3: Illustration of `DreamLIP` with synthetic images generated by SDXL-turbo.

Table 2: Zero-shot transfer evaluation of different models. Performance on ImageNet and 10 common downstream datasets are reported.

Data	Model	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	Average	ImageNet
Model Architecture: ViT-B/32													
CC3M	CLIP	10.2	71.3	32.1	33.8	1.4	1.0	12.0	12.1	50.9	10.8	23.6	17.2
	DreamLIP	16.1	82.0	45.4	41.3	2.5	1.0	13.9	18.8	64.4	14.1	**30.0**	**25.9**
	+ Synthetic Images	21.7	80.9	51.2	45.1	3.2	1.5	20.8	24.1	68.2	15.6	**33.2**	**29.4**
CC12M	CLIP	26.5	72.5	38.0	37.1	13.7	2.6	11.4	46.2	74.0	25.7	34.8	32.9
	DreamLIP	48.9	86.4	63.0	55.7	17.9	1.9	23.5	41.9	83.2	25.8	**44.8**	**44.2**
	+ Synthetic Images	46.0	86.1	57.2	53.3	26.1	3.3	26.3	56.4	83.7	31.3	**47.0**	**46.4**
YFCC15M	CLIP	26.9	77.8	48.2	42.5	5.5	4.7	18.5	15.7	62.0	39.0	34.1	33.3
	DreamLIP	51.7	87.9	60.7	54.8	9.4	7.1	26.8	36.3	79.6	48.6	**46.3**	**46.6**
	+ Synthetic Images	54.2	87.1	57.9	53.5	14.1	8.7	31.4	32.3	80.8	40.5	**46.1**	**47.5**
Model Architecture: ViT-B/16													
CC3M	CLIP	10.3	54.9	21.8	25.0	0.8	1.4	10.5	12.8	43.3	10.2	19.1	20.3
	DreamLIP	19.4	74.3	44.2	45.9	2.8	1.0	17.0	27.1	63.1	14.7	**31.0**	**31.1**
	+ Synthetic Images	22.8	72.8	43.0	46.6	3.9	1.2	22.1	25.7	70.0	17.7	**32.6**	**33.3**
CC12M	CLIP	25.3	66.5	32.1	39.9	14.7	1.9	13.5	45.0	59.8	15.0	31.4	34.0
	DreamLIP	58.3	87.3	62.6	54.3	29.7	4.9	29.2	60.3	83.1	28.9	**49.9**	**50.3**
	+ Synthetic Images	58.3	87.3	64.6	53.9	29.7	4.9	29.2	60.3	83.1	28.9	**50.0**	**50.7**
YFCC15M	CLIP	35.0	67.1	34.8	42.0	5.1	6.3	13.9	20.4	54.5	44.3	32.3	34.1
	DreamLIP	44.2	89.0	62.0	57.1	9.2	6.4	30.5	32.6	79.8	40.2	**45.1**	**48.2**

D Evaluation on Vision-Language Compositionality

We compare DreamLIP with CLIP on Attribution, relation, ordering (ARO) [11] and SugarCrepe [7] benchmark. These two benchmarks are used to measure compositional understanding of vision-language models. Results are shown in Table 3. DreamLIP significantly outperforms CLIP across all tasks when pretrained on the same dataset (Merged-30M). Also, DreamLIP-30M achieves better results on 7 out of 11 tasks compared to CLIP-400M. It indicates the usage of long captions with detailed descriptions well enhance model's compositional understanding.

Table 3: Results on the ARO [11] and SugarCrepe [7] benchmark. CLIP-30M and -400M indicate the CLIP is pre-trained on Merged-30M and LAION-400M dataset, respectively. DreamLIP-30M is pre-trained on Merged-30M dataset

Model	**Aro**				**SugarCrepe**						
	VG		COCO	Flickr	Replace			Swap		Add	
	Attribution	Relation	Order	Order	Object	Attribute	Relation	Object	Attribute	Object	Attribute
CLIP-30M	58.28	43.52	23.42	27.02	82.99	73.10	59.25	60.00	65.17	69.84	63.15
DreamLIP-30M	**78.17**	**53.62**	**41.26**	**42.78**	91.46	82.23	**72.40**	**69.80**	**79.43**	81.38	77.46
CLIP-400M	59.92	47.86	38.83	42.56	**91.58**	**82.75**	67.57	61.22	68.62	**82.01**	**78.61**

E Experiments

E.1 Hyper-Parameters

Table 4 provides an overview of the pre-training hyperparameters used for CLIP on all datasets. Further details can be found in Table 4. The pre-training processes of CC12M, YFCC15M and Merged-30M were conducted on four machines with eight A100 GPUs. For CC3M, eight A100 GPUs are used.

Table 4: Detailed pre-training hyper-parameters for CLIP training on all four image-text datasets.

Config	Value
Batch size	$1,024$
Optimizer	AdamW [8]
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(a) Hyper-parameter on CC3M.

Config	Value
Batch size	$8,192$
Optimizer	AdamW [8]
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(b) Hyper-parameter on CC12M.

Config	Value
Batch size	$8,192$
Optimizer	AdamW [8]
Learning rate	5×10^{-4}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(c) Hyper-parameter on YFCC15M.

Config	Value
Batch size	$8,192$
Optimizer	AdamW [8]
Learning rate	5×10^{-4}
Weight decay	0.2
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-6}
Total epochs	32
Warm up iterations	$2,000$
Learning rate schedule	cosine decay

(d) Hyper-parameter on Merged-30M.

E.2 Additional Ablation Study

Different Image Backbones in Semantic Segmentation. Table 5 shows the transferable performance of CLIP and DreamLIP with different image backbones on semantic segmentation tasks. DreamLIP always achieves better performance than CLIP across different pre-training data and different image backbones.

Ablation study of σ. Table 6 shows the performance of DreamLIP when adjusting σ. σ controls the sparsity of subcaption-specific grouping visual tokens,

Table 5: Transferable performance of semantic segmentation on ADE-847, PC-459, ADE-150, PC-59, and VOC-20. Following SAN [10], we used the full training set of COCO-stuff as the training data and our `DreamLIP` as pretrained models.

Data	Method	ADE-847	PC-459	ADE-150	PC-59	VOC-20	avg.
Model Architecture: ViT-B/32							
CC3M	CLIP	2.1	5.2	12.3	33.8	65.4	23.8
	`DreamLIP`	4.1	7.5	17.1	39.9	76.5	29.0
CC12M	CLIP	3.3	6.7	15.7	39.2	79.7	28.9
	`DreamLIP`	6.1	10.0	23.3	43.6	85.5	33.7
YFCC15M	CLIP	3.2	8.1	14.4	42.0	82.3	30.0
	`DreamLIP`	6.4	11.1	22.4	48.9	88.2	35.4
Merged-30M	CLIP	5.8	10.2	21.0	45.8	86.9	33.9
	`DreamLIP`	**8.1**	**12.5**	**25.3**	**49.9**	**90.9**	**37.3**
Laion-400M	CLIP	6.1	12.2	21.3	46.3	88.3	34.8
Model Architecture: ViT-B/16							
CC3M	CLIP	1.9	5.3	11.4	34.5	64.4	23.5
	`DreamLIP`	4.9	8.7	20.5	45.0	84.5	32.7
CC12M	CLIP	3.4	7.9	16.4	39.5	80.4	29.5
	`DreamLIP`	6.6	12.3	23.7	48.4	85.2	35.2
YFCC15M	CLIP	1.2	4.9	13.9	41.5	74.0	27.1
	`DreamLIP`	6.6	13.5	24.7	51.4	90.9	37.4
Merged-30M	CLIP	7.3	12.1	25.6	49.1	86.4	36.1
	`DreamLIP`	9.8	**15.4**	**30.6**	**55.1**	92.2	**40.6**
Laion-400M	CLIP	**10.1**	12.6	27.5	53.8	**94.0**	**40.6**

as shown in Eq.6 in the main paper. Larger σ results in sparser subcaption-specific grouping visual tokens within an image.

Table 6: Ablation study of σ. σ is a sparsity threshold as shown in Eq.6 in the main paper. It controls the sparsity of subcaption-specific grouping visual tokens. ViT-B/16 is used as the image backbone. All models are pretrained on CC3M.

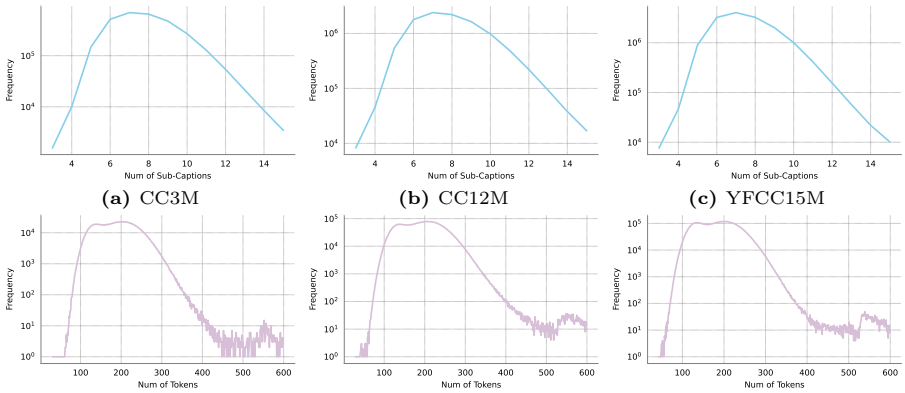
	Text Retrieval		Image Retrieval		Classification	Segmentation
σ	Flickr30k	MSCOCO	Flickr30k	MSCOCO	ImageNet	VOC-20
	R@1	R@1	R@1	R@1	Acc.(%)	mIOU
CLIP	29.4	14.3	19.9	10.2	16.0	62.7
CLIP*	32.6	14.8	21.4	11.5	20.3	64.4
0.0	69.5	42.8	54.4	30.4	31.1	84.5
0.1	71.0	41.2	54.3	30.4	31.9	84.2
0.3	71.1	41.6	54.2	30.3	31.7	84.0
0.5	70.9	42.4	54.4	30.0	31.7	85.0
0.7	72.5	42.2	54.9	30.4	31.8	85.8

Ablation study of λ_S. We present the influence of λ_S in Table 7. λ_S denotes the weight of fine-grained alignment contrastive loss, as shown in Eq.9 in the main paper. For most global-level tasks (retrieval and classification), the best performance is reached when $\lambda_S = 0.7$. For local-level tasks, *i.e.*, semantic segmentation, the best performance is reached when $\lambda_S = 0.9$. The results

indicate the fine-grained alignment contrastive loss is helpful for vision-language alignment. Larger λ_S leads the model to learn more fine-grained clues.

Table 7: Ablation study of λ_S, which controls the weight of fine-grained alignment contrastive loss. ViT-B/16 is used as image backbone. All models are pretrained on CC3M.

λ_S	Text Retrieval		Image Retrieval		Classification	Segmentation
	Flickr30k	MSCOCO	Flickr30k	MSCOCO	ImageNet	VOC-20
	R@1	R@1	R@1	R@1	Acc.(%)	mIOU
CLIP	29.4	14.3	19.9	10.2	16.0	62.7
CLIP*	32.6	14.8	21.4	11.5	20.3	64.4
0.1	69.5	42.8	54.4	30.4	31.1	84.5
0.3	71.0	41.2	54.3	30.4	31.9	84.2
0.5	70.2	42.2	55.0	30.8	31.7	83.6
0.7	71.4	42.6	55.6	31.6	32.1	84.6
0.9	71.1	42.2	55.9	31.5	31.9	85.0



(a) CC3M **(b)** CC12M **(c)** YFCC15M

(d) CC3M **(e)** CC12M **(f)** YFCC15M

Fig. 4: Some statistics of long captions generated by ShareGPT4V. (a)-(c) refer to number of sub-captions from long captions; (d)-(f) refer to number of tokens in long captions.

E.3 Statistic of Long Captions on Different Datasets

We conducted some statistics (*i.e.*, the number of tokens and sub-captions) of long captions generated by ShareGPT4V on different datasets. (a)-(c) refer to number of sub-captions from long captions; (d)-(f) refer to number of tokens in long captions.

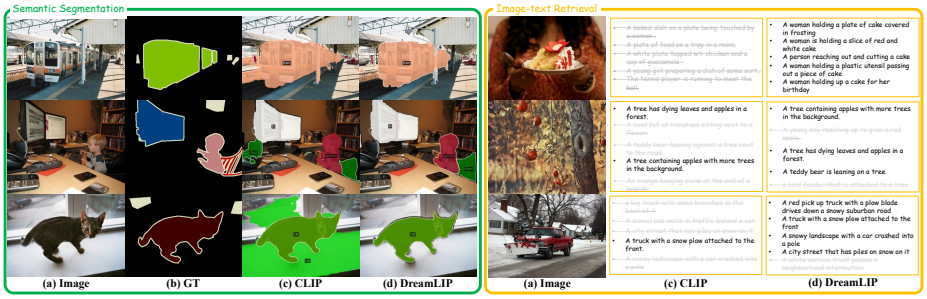


Fig. 5: Visualization of semantic segmentation and image-text retrieval.

F Visualization

F.1 Visualization of Semantic Segmentation and Image-text Retrieval.

In order to offer a comprehensive qualitative understanding, we have curated a set of examples from the VOC and MSCOCO validation sets in Figure 5, showcasing the notable accuracy improvements achieved by `DreamLIP`. These carefully chosen examples serve as compelling evidence of our method's remarkable ability to effectively distinguish between intricate and nuanced categories. Notably, our method excels in scenarios where vanilla CLIP encounters challenges and struggles to make accurate differentiations. By presenting these examples, we substantiate the claim that our approach significantly enhances the discriminative power of the model, particularly in fine-grained categorization tasks.

F.2 Attention Map Visualization.

We visualize the attention map between different sub-captions from the generated long captions in Fig.6 following [2, 3]. As we motivate above, `DreamLIP` can indeed focus on the corresponding regions according to the different sub-captions. Even the dog's tongue (as shown in row 2 and column 2) and the microphone (as shown in row 3 and column 3) in the noisy background can be precisely perceived by `DreamLIP`.

G Limitations

The existing multimodal large models suffer from hallucinations, with longer captions leading to more severe hallucinations. Directly using the generated long captions will introduce much noise. How to solve the multimodal hallucination problem under long captions, which can further improve the performance of our method.

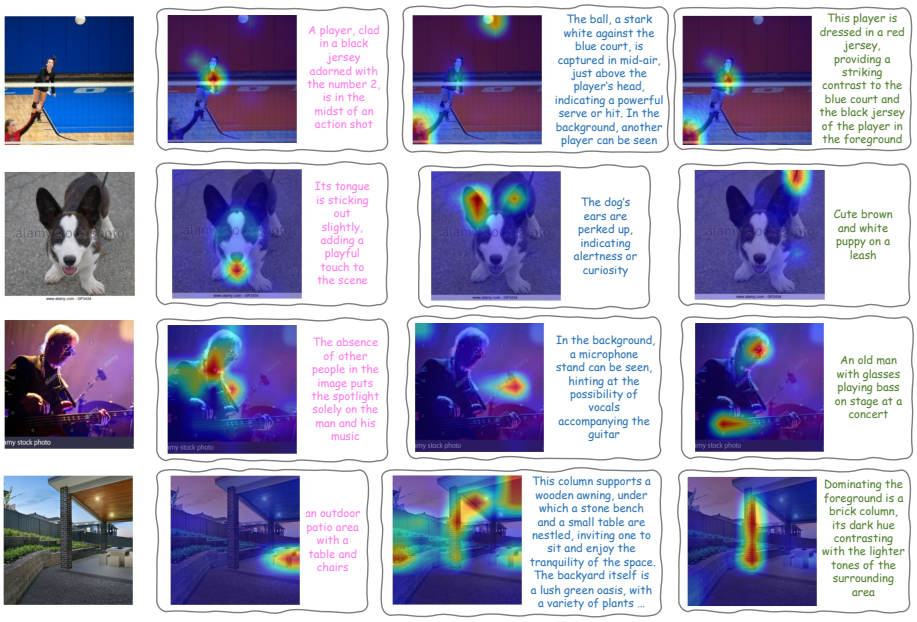


Fig. 6: Visualization for Attention Map. The sub-captions corresponding to the attention maps are split from the generated long captions.

References

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
2. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Int. Conf. Comput. Vis. pp. 397–406 (2021)
3. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 782–791 (2021)
4. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. arXiv preprint arXiv:2403.04692 (2024)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn. (2020)
7. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Adv. Neural Inform. Process. Syst. **36** (2024)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv:1711.05101 (2017)

9. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023)
10. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2945–2954 (2023)
11. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: Int. Conf. Learn. Represent. (2022)